Text Categorisation on Semantic Analysis for Document Categorisation Using a World Knowledge Ontology

Xiaohui Tao*, Patrick Delaney and Yuefeng Li

E-mail: {Xiaohui.Tao, Patrick.Delaney}@usq.edu.au, y2.li@qut.edu.au

***Abstract*—An effective text categorisation approach can allow users easy access to useful and meaningful textual information. However, while many automatic categorisation techniques have been developed, there is still room for improvement in categorisation performance. In this work, we have proposed an innovative approach using a large world knowledge ontology built from the Library of Congress Subject Headings (LCSH) to categorise text documents. The semantic content of documents is represented by well-defined and well-specified subjects extracted from the ontology. The proposed approach has been successfully evaluated, using a large data set with linguist-generated categorisation results in empirical experiments.**

I. Introduction

An effective categorisation method can improve the efficiency of systems in accessing textual information. In particular, Web personalization systems benefit from categorising a user's local documents (e.g. browsing history, emails, tweets, and blogs) to concepts in a global knowledge base [37], [34]. A user profile is the simulation of the user's concept model, whereas a user's concept model is the user's local reflection of world knowledge with only the topics of interest to the user [35]. User local documents provide wealthy user background knowledge. Therefore, to acquire quality user profiles, user background knowledge needs to be discovered from user local documents and global world knowledge. Figure 1 illustrates a scenario of user profile acquisition, which is completed by discovering user background knowledge from the categorisation of user local documents to subjects in a world knowledge ontology. This method was successfully accomplished and evaluated by Tao et al. [35], using a world ontology constructed from the Library of Congress Subject Headings (LCSH) [1]. Other successfully accomplished models include Sieg et al. [32] using the Open Directory Project [2].

Effective document categorisation is mostly completed by human effort, with well-trained experts (e.g. linguists, librarians, and metadata experts) manually categorising documents in either traditional or digital forms and assigning descriptors (a list of subjects) to documents [9]. However, manual categorisation is expensive and time-consuming. Additionally, manual categorisation becomes problematic when dealing with large repositories. Many automatic methods have been developed to categorise documents based on semantic contents (e.g. [7], [1], [4], [21], [26], [38], [39]). However, in pursuing efficient, automatic categorisation, two problems were revealed: (i) knowledge bases chosen for categorisation are usually inadequate for describing the real world, being either constructed in over-simplified structures or covering only a limited range of topics, resulting in inadequate subjects being assigned to documents for categorisation; and (ii) imperfect categorisation algorithms still have large room for improvement. Inadequate subjects have been used to categorise documents because imperfect algorithms were used. Kasper et al. [18] reviewed implicit, explicit and hybrid user acquisition frameworks, noting that there is a clear need to investigate the potential of learning algorithms. Refining user acquisition models against these identified problems will improve semantic content based document categorisation and add to research on user background knowledge discovery.

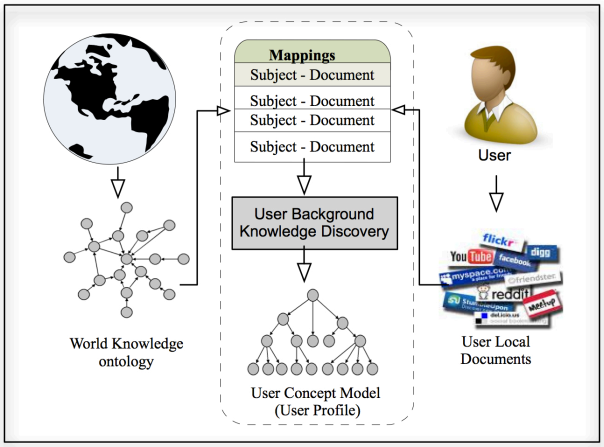


Fig. 1. Application of Document Categorisation exploiting World Knowledge

In this paper, we propose an automatic semantic categorisation approach using the LCSH ontology, a mature and well-defined world knowledge ontology previously evaluated by Tao et al. [35] successfully. Given a document, its semantic features are first discovered. The document's relevant subjects are then extracted from the world ontology based on the discovered features. Finally, these subjects are generalised in order to assign a list of competent subjects to the document for categorisation. An empirical experiment evaluated the proposed approach by comparing it against typical categorisation methods including *Rocchio* and kNN, based on the ground

[1] http://id.loc.gov/authorities/

[2] http://www.dmoz.org/

truth of manual categorisation results. A large corpus was collected from the real world for the experiments. The evaluation result was promising and encouraging. The proposed approach makes the following contributions to research and practice:

- A method that categorises documents into multiple categories based on semantic analysis of content;
- An innovative algorithm that generalises subjects based on their semantic relationships;
- A novel approach using a large world knowledge ontology to guide semantic categorisation of documents.

Semantic categorisation may also help capture users' demands and opinions in e-Commerce, as well as the rivals' intelligence, benefit from the improved efficient access of public text documents.

The paper is organised as follows: Section II discusses the related work; Section III formalises the definitions and the research problem in this work. After that, Section IV introduces the proposed semantic categorisation method. The experiment design is described in Section VI and the results are discussed in Section VII. Finally, Section VIII makes the conclusions.

II. Related Work

The semantic content of text documents has different representations, such as lexicons, categories, or patterns. A lexicon-based representation of documents is easily understood by users and computational systems. Text documents are represented by a set of descriptors chosen from controlled vocabularies defined in terminological ontologies, thesauruses, or dictionaries. However, when extracting lexical descriptors, some noisy descriptors are also extracted alongside meaningful, representative descriptors, due to the term ambiguity problem. The development of terminological ontologies, thesauruses, or dictionaries is also financially expensive and time-consuming, due to the large requirement of human effort. As a result, the lexicon-based representation of semantic content is inefficient.

Categorisations are widely used in methods to represent document contents, like those in [33], [29], [14], [35]. In this approach, the concepts revealed from text are represented by categories and organised in a tree or graphic structure. The relationships existing between concept nodes in the structure are explored in order to measure the competency of a concept describing or representing the content. However, usually simple relations (subsumption of one containing another or super- and sub-class) are used in categorisations rather than detailed, well-specified semantic relations (like is-a, part-of, and related-to). Thus, categorisation-based representation needs to improve for more detailed and precise levels of concept specification.

Pattern-based representation uses multiple phrases to represent document content [12], [10], [22], [24]. However, pattern-based categorisation suffers from issues caused by the length of patterns. Concepts are specific and discriminating only with substantially long patterns, but long patterns have low frequency. Consequently, the power of long patterns reduces because low frequency makes the patterns less applicable to problems [23]. In addition, because of the text-mining techniques used for pattern discovery, sometimes noisy patterns are extracted alongside useful patterns. Alternative weighting methods need to be investigated to overcome this problem in pattern-based content representation.

Many works utilise pattern-mining techniques to help build classification models, which is similar as the strategy employed in our work. Malik and Kender [28] proposed the "Democratic Classifier", a pattern-based classification algorithm using short patterns. However, the democratic classifier relies on the quality of training samples and cannot deal with the "no training set available" problem. Bekkerman and Matan [3] argued that most of the information on documents can be captured in phrases, and they proposed a text classification method that employs lazy learning from labelled phrases. The phrases in their work are in fact a special form of sequential patterns that are used in our work for feature extraction of documents.

Text classification is a common technique used to classify a stream of documents into categories by using the classifiers learned from the training samples [25]. This can be two types: *kernel-based* and *instance-based* [2]. Typical kernel-based classifier learning approaches include *Support Vector Machines* (SVMs) [17] and regression models [31]. Kernel-based approaches sometimes incorrectly classify negative samples into positive. Typical instance-based classification approaches include the kNN and its variants, which do not rely upon the statistical distribution of training samples. However, the instance-based approaches become unstable when classifying highly accurate positive samples from an unlabelled data set. Other reports, such as [30], have a different view and categorise text classification techniques into *document representations based classifiers* including SVMs and kNN and *word probabilities based classifiers* including Naive Bayesian, decision trees [17] and neural networks [44]. These classification techniques have different strengths and weaknesses, and should be chosen carefully depending on the problem space.

Unsupervised text classification aims to classify documents into classes that are absent of any labelled training documents. Many successful models have been proposed, such as [43]. However, on many occasions, the target classes may not have any labelled training documents available. One particular example is the "cold start" problem in recommender systems and social tagging [13]. Unsupervised classification can automatically learn an annotation model to make recommendations or label the tags when the products or tags are rare and have no useful associated information. Without associated training samples, Yang et al. [42] built a classification model for a target class by analysing the correlating auxiliary classes. The work in this paper is similar to that model, however, our model differs by exploiting a hierarchical world knowledge ontology for classification, instead of only auxiliary classes. Also exploiting a world knowledge base, Yan et al. [40] examined unsupervised relation extraction from Wikipedia articles and integrated linguistic analysis with web frequency information to improve unsupervised classification performance. By comparison, our work aims to exploit a world knowledge ontology to help unsupervised classification. Cai et al. [6] and Houle and Grira [16] proposed unsupervised approaches to evaluate and improve the quality of selecting features. Given a set of data, their approach finds a subset containing the most informative, discriminative features. Though the work presented in this paper also relies on features selected from documents, the features are further investigated with their referring-to ontological concepts to improve the performance of classification.

Ontologies have been used to facilitate text classification by generating features using domain-specific and common-sense knowledge in large ontologies [11] and semantic relations in web personalization [34] and document retrieval [27]. Camous et al. [8] introduced a domain-independent method that uses the Medical Subject Headings (MeSH) ontology. The method observes the inter-concept relationships and represents documents by MeSH subjects, considering semantic relations. Another world ontology commonly used in text classification is Wikipedia [1]. For instance, Hu et al. [14] derived background knowledge from Wikipedia to represent documents and attempted to deal with the sparsity and high dimensionality problems in text classification. Compared to this prior research, our work uses the LCSH, a superior world knowledge ontology under continuous development for a hundred years by knowledge engineers.

Text classification models were originally designed to handle only single-label problems, where each document is classified into only one class. However, in many circumstances single-label text classification is inadequate, such as with social networks where multiple labels are needed [15], [20]. Similar to the work of Yang et al. [41], our method also targets multi-label text classification. However, rather than adopting active learning algorithms for multi-label classification, we exploit concepts and their structure in world knowledge ontologies [19].

III. Research Problem and Definitions

Let $\mathcal{D} = \{d_i \in \mathbb{D}, i = 1, \ldots, m\}$ be a set of text documents; $\mathcal{S} = \{s_1, \ldots, s_K\}$ be a large set of classes, where K is the number of classes. If there is an available training set $\mathcal{D}_t = \{d_j \in \mathbb{D}, j = m+1, \ldots, n\}$ with $y_j^k = \{0, 1\}, k = 1, \ldots, K$ provided for describing the likelihood of d_j belonging to class s_k, it is easy to learn a binary prediction function $p(y^k|d)$ and use it to classify $d_i \in \mathcal{D}$. However, our objective is to learn a prediction function $p(y^k|d)$ to classify d_i into $\{s_k\} \subset \mathcal{S}$ without $\mathcal{D}_t$ available. We refer to this problem as *unsupervised multi-label text classification*.

***Definition** 1:* Let $\Omega = \{d_1, d_2, d_3, \ldots, d_n\}$ be a finite and non-empty set of text documents. Given $d \in \Omega$, its semantic content can be categorised by using the mapping:

$$\eta : \Omega \to 2^{\mathcal{S}}, \quad \eta(d) = \{s \in \mathcal{S} | str(d, s) \geq min\_str\} \subseteq \mathbb{S}$$

and its reverse mapping:

$$\eta^{-1} : \mathcal{S} \to 2^{\Omega}, \quad \eta^{-1}(s) = \{d \in \Omega | str(d, s) \geq min\_str\} \square$$

Note that $str(d, s)$ is the strength describing the competency of s to categorise d, and min_str is the threshold defining the desirable competency level.

To illustrate the problem, a sample document is shown in Fig. 2. This screenshot was taken from the online catalogue of the University of Melbourne Library[3]. The catalogue information is about a book with the title and summarised content:

> *Economic espionage and industrial spying. Dimensions of economic espionage and the criminalization of trade secret theft – Transition to an information society - increasing interconnections and interdependence – International dimensions of business and commerce – Competitiveness and legal collection versus espionage and economic crime – Tensions between security and openness – The new rule for keeping secrets - the Economic Espionage Act – Multinational conspiracy or natural evolution of market economy.*

and a list of librarian manually-assigned subjects:

> *Business intelligence; Trade secrets; Computer crimes; Intellectual property; Commercial crimes.*

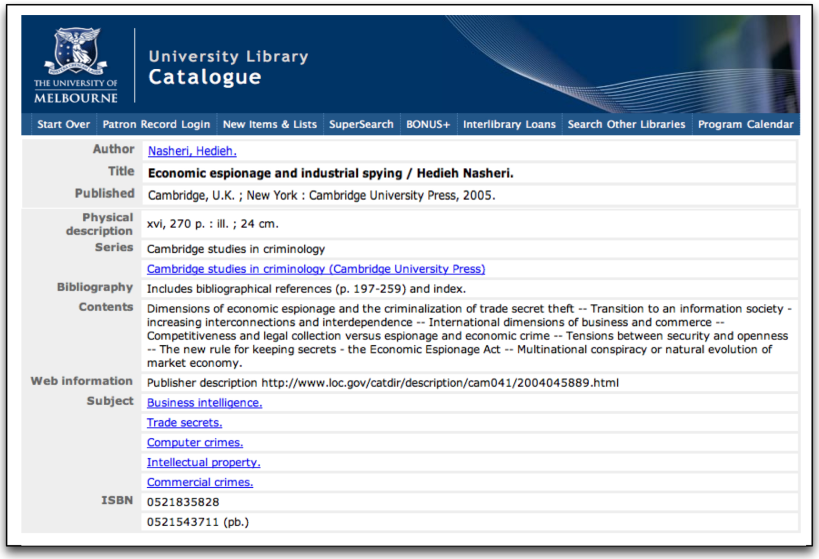
The title, summarised content, and subjects in Fig. 2 depict the ultimate goal we pursue: given a text document (e.g., the title and summarised content in Fig. 2), categorise it to an indexed set of subjects extracted from the world ontology (e.g., the listed subjects in Fig. 2). Ideally, the extracted subjects should be the same as these linguist manually-assigned subjects, because they represent human intellectual work in semantic categorisation. However, at this stage, attaining the same result as human work is unrealistic. Therefore, finding similar assignment of subjects with human work is the aim of our work. The sample case in Fig. 2 will be used through the rest of the paper to assist the explanation.

The world knowledge ontology is constructed from the Library of Congress Subject Headings (LCSH), a knowledge system developed for organising information in large library collections. It has been under continuous development for over a hundred years to describe and classify human knowledge. Because of the dedicated endeavours of knowledge engineers from generation to generation, the LCSH has become a de facto standard for concept cataloguing and indexing, superior to other knowledge bases. Tao et al. [35] previously compared the LCSH with the Library of Congress Classification, the Dewey Decimal Classification, and Yahoo! categorisation, and reported that the LCSH has broader topic coverage, more meaningful structure, and more accurate semantic relations. The LCSH has been widely used as a means for many knowledge engineering and management works [9]. In this work, the class set $\mathcal{S} = \{s_1, \ldots, s_K\}$ is encoded from the LCSH subject headings.

***Definition** 2: (SUBJECT)* Let $\mathcal{S}$ be the set of subjects, an element $s \in \mathcal{S}$ is a 4-tuple $s := \langle label, neighbour, ancestor, descendant \rangle$, where

- label is a set of sequential terms describing s; $lable(s) = \{t_1, t_2, \ldots, t_n\}$;
- $neighbour$ refers to the set of subjects in the LCSH that directly link to s, $neighbour(s) \subset \mathcal{S}$;

[3] http://cat.lib.unimelb.edu.au/. Note that the screenshot has been altered for display - the alteration was completed without pruning away any meaningful content.



Author	Nasheri, Hedieh.
Title	**Economic espionage and industrial spying / Hedieh Nasheri.**
Published	Cambridge, U.K. ; New York : Cambridge University Press, 2005.
Physical description	xvi, 270 p. : ill. ; 24 cm.
Series	Cambridge studies in criminology
	Cambridge studies in criminology (Cambridge University Press)
Bibliography	Includes bibliographical references (p. 197-259) and index.
Contents	Dimensions of economic espionage and the criminalization of trade secret theft -- Transition to an information society - increasing interconnections and interdependence -- International dimensions of business and commerce -- Competitiveness and legal collection versus espionage and economic crime -- Tensions between security and openness -- The new rule for keeping secrets - the Economic Espionage Act -- Multinational conspiracy or natural evolution of market economy.
Web information	Publisher description http://www.loc.gov/catdir/description/cam041/2004045889.html
Subject	Business intelligence.
	Trade secrets.
	Computer crimes.
	Intellectual property.
	Commercial crimes.
ISBN	0521835828
	0521543711 (pb.)

Fig. 2. A Sample Document with Subjects Manually Assigned by Librarians.

- $ancestor$ refers to the set of subjects directly and indirectly link to s and locating at more abstractive level than s in the LCSH, $ancestor(s) \subset \mathcal{S}$;
- $descendant$ refers to the set of subjects directly and indirectly link to s and locating at more specific level than s in the LCSH, $descendant(s) \subset \mathcal{S}$. □

The semantic relationships of subjects are encoded from the references defined in the LCSH for subject headings, including *Broader Term*, *Used for*, and *Related to*. The $ancestor(s)$ in Definition 2 returns the *Broader Term* subjects of s; the $descendant(s)$ is the reversed function of $ancestor(s)$, with additional subjects *Used for* s; the $neighbour(s)$ returns the subjects *Related to* s.

With Definition 2, the world knowledge ontology is defined:

***Definition** 3: (ONTOLOGY)* Let $\mathcal{O}$ be a world ontology. $\mathcal{O}$ contains a set of subjects linked by their semantic relations in a hierarchical structure. $\mathcal{O}$ is a 3-tuple $\mathcal{O} := \langle \mathcal{S}, \mathcal{R}, \mathcal{H}^{\mathcal{S}}_{\mathcal{R}} \rangle$, where

- $\mathcal{S}$ is the set of subjects defined in Definition 2;
- $\mathcal{R}$ is the set of relations linking any pair of subjects;
- $\mathcal{H}^{\mathcal{S}}_{\mathcal{R}}$ is the hierarchical structure of $\mathcal{O}$ constructed by $\mathcal{S} \times \mathcal{R}$. □

IV. Theoretical Framework

A lexicon-based representation is based on the statistic of occurring terms. Such a representation is easy to understand by users and systems. However, along with meaningful, representative features, some noisy terms are also extracted, caused by sense ambiguity of terms. To deal with this problem, pattern-based representation is studied, which uses frequent sequential patterns (phrases) to represent document contents [24]. The pattern-based representation is superior to lexicon-based, as the context of terms co-occurred in phrases is considered. However, the pattern-based presentation suffers from a limitation caused by the length of patterns. Though a long pattern is wealthy with information and so more discriminative, it usually has low frequency and as a result, becomes inapplicable. To overcome the problem, we represent the content of documents by a set of weighted closed frequent sequential patterns discovered by pattern mining techniques.

***Definition** 4: (FEATURES)* Given a document $d = \{t_1, t_2, \ldots, t_n\}$ as a sequential set of repeatable terms, the feature set, denoted as $\mathcal{F}(d)$, is a set of weighted phrase patterns, $\{\langle p, w(p) \rangle\}$, extracted from d that satisfies the following constraints:

- $\forall p \in \mathcal{F}(d), p \subseteq d$.
- $\forall p_1, p_2 \in \mathcal{F}(d)(p_1 \neq p_2), p_1 \not\subset p_2 \wedge p_2 \not\subset p_1$.
- $\forall p \in \mathcal{F}(d), w(p) \geqslant \vartheta$, a threshold. □

The initial classification of d to $s_k \in \mathcal{S}$ is done through accessing a term-subject matrix created by the subjects and their labels. Adopting the features discovered previously, we use a feature-subject mapping approach to initially assign subject classes to the document.

***Definition** 5: (TERM-SUBJECT MATRIX)* Let $\mathcal{T}$ be the term space of $\mathcal{S}, \mathcal{T} = \{t \in \bigcup_{s \in \mathcal{S}} label(s)\}, \langle \mathcal{S}, \mathcal{T} \rangle$ is the matrix coordinated by $\mathcal{T}$ and $\mathcal{S}$, where a mapping exists:

$$\mu : \mathcal{T} \rightarrow 2^{\mathcal{S}}, \quad \mu(t) = \{s \in \mathcal{S} | t \in label(s)\}$$

and its reverse mapping also exists:

$$\mu^{-1} : \mathcal{S} \rightarrow 2^{\mathcal{T}}, \quad \mu^{-1}(s) = \{t \in \mathcal{T} | s \in \mu(t)\} \qquad \square$$

Adopting Definition 4 and 7, we can initially classify $d_i \in \mathcal{D}$ into a set of subjects using the following prediction:

$$\widehat{y_i^k} = I(s_k \in h \circ g \circ f(d_i)), i = 1, \ldots, m \qquad (1)$$

where $I(z)$ is an indicator function that outputs 1 if z is true and zero, otherwise; $f(d) = \{p | \langle p, w(p) \rangle \in F(d)\}$; $g(\rho) = \{t \in \cup_{p \in \rho} p\}$; $h(\tau) = \{s \in \cup_{t \in \tau} \mu(t)\}$.

The initial classification process easily generates noisy subjects because of direct feature-subject mapping. Against the problem, we introduce a method to generalise the initial subjects to optimise the classification. We observed that in initial classification some subjects extracted from the ontology are overlapping in their semantic space. Thus, we can optimise the classification result by keeping only the dominating subjects and pruning away those being dominated. This can be done by investigating the semantic relations existing between subjects. Let s_1 and s_2 be two subjects and $s_1 \in ancestor(s_2)$ ($s_2 \in descendant(s_1)$). s_1 refers to an broader semantic space than s_2 and thus, is more general. Vice versa, s_2 is more specific and focused than s_1. Hence, if some subjects are covered by a common ancestor, they can be replaced by the common ancestor without information loss. The common ancestor is unnecessary to be chosen from the initial classification result, as choosing an external common ancestor also satisfies the above rule. After generalising the initial classification result, we have a smaller set of subject classes, with no information lost but some focus. (The handling of focus problem is presented in next section.)

***Definition** 6: (GENERALISED CLASSIFICATION)* Given a document d and its initial classification result, a subject set denoted by $S^I(d)$, the generalised classification result, denoted as $S^G(d)$, is the set of subjects satisfying:

1) $\forall s \in S^I(d), \exists s' \in S^G(d), s \neq s', s \in descendants(s')$.
2) $\forall s_1, s_2 \in S^G(d)(s_1 \neq s_2), s_1 \notin descendants(s_2) \wedge s_2 \notin descendants(s_1)$.

V. Framework in Practice

To design a semantic content-based document categorisation approach, two critical difficulties must be addressed: choosing a competent knowledge base, and proposing a categorising algorithm with less imperfection. This work was designed to address these two difficulties. A world knowledge ontology constructed from the LCSH is utilised to work as the knowledge base for the semantic content based categorisation. Documents are categorised to the subjects in the LCSH ontology through three steps: discovering features from the documents; extracting subjects from the LSCH ontology based on the features; generalising the subjects to finalise categorisation. The conceptual framework for the design is illustrated in Fig. 3, which consists of three modules, each one designed for one step.

Feature Discovery Module. Pattern Taxonomy Method has been employed in this module to discover features from the given document, based on the theory of closed frequent sequential patterns. As the outcome of this module, a set of patterns with weights greater than a minimum value is selected to represent the features of the document;

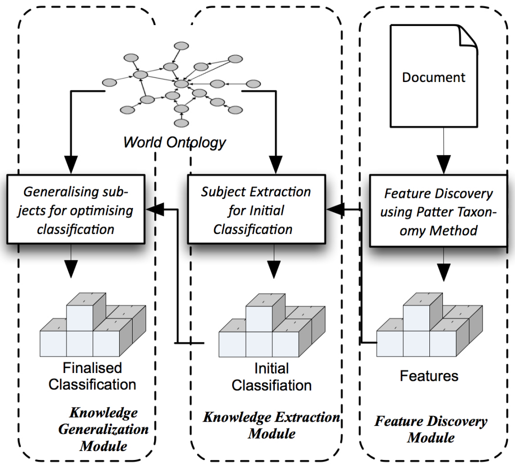

Document
World Ontplogy
Generalising sub-jects for optimising classification
Subject Extraction for Initial Classification
Feature Discovery using Patter Taxon-omy Method
Finalised Classification
Initial Classifiation
Features
Knowledge Generalization Module
Knowledge Extraction Module
Feature Discovery Module

Fig. 3. Conceptual Framework

Knowledge Extraction Module. A term-subject matrix has been established in this module to extract appropriate subjects from the LCSH world ontology, based on the features extracted in previous step. The matrix has two attributes: joint set of terms from the label of all subjects; the set of all subjects in the world ontology. Given a set of patterns (features), a mapping set of subjects is extracted, in which each element is assigned with a strength value representing its level of competency to categorise the document;

Knowledge Generalisation Module. The subjects extracted in previous step are investigated in this module for their semantic relations with other subjects in the neighbourhood and their location in the structure of the world ontology. The subjects referring to common semantic space are merged and replaced by their common ancestor subject. Finally, a refined indexed list of subjects are generalised to represent the semantic content and to categorise the document.

The LCSH world ontology and proposed semantic categorisation approach is be explained in the following sections.

A. The LCSH World Ontology

Textual information has some properties that make semantic categorisation difficult. The structure and format of text documents are usually complex and the topics are heterogeneous, meaning the content may change constantly [36]. An efficient text document categorisation method must deal with these properties. As shown in many previous works like [46], [35] and [32], an effective strategy is using world knowledge ontologies. Ontologies are formal descriptions and specifications of conceptualisation. By nature, ontologies are a powerful technique for clarifying and then solving complex, heterogeneous problems. World knowledge is commonsense knowledge possessed by people and acquired through their experiences and education [45]. To categorise text documents with constant changes, world knowledge provides constant support because it updates alongside the progress of civilisation. The ontology (or any knowledge base) chosen to guide efficient, automatic text categorisation should be competent to deal with these properties.

The world knowledge ontology in this work is constructed based on the LCSH, similar to the work of [35]. The LCSH was developed for organising and retrieving information from a large volume of library collections. As discussed by Chan [9], the LCSH has many superiorities for handling the problems in text categorisation:

- The LCSH system is an ideal world knowledge base covering an exhaustive range of topics (Competent to deal with the complexity and heterogeneity problems);
- The LCSH represents the natural growth and distribution of human intellectual work. For over a hundred years, the knowledge contained in the LCSH has undergone continuous revision and enrichment (Competent to deal with the constant change problem);

TABLE I
COMPARISON OF DIFFERENT WORLD TAXONOMIES

	LCSH	**LCC**	**DDC**	**Yahoo!**
# of topics	394,070	4,214	18,462	100,000
Structure	Directed Acyclic Graph	Tree	Tree	Directed Acyclic Graph
Depth	37	7	23	10
Semantic Relations	Broader, Used-for, Related-to	Super- and Sub-class	Super- and Sub-class	Super- and Sub-class

- The LCSH has the most comprehensive non-specialised controlled vocabulary in English (Providing competent subjects to categorise documents.)

Though the majority of libraries utilising the LCSH are located in the United States, almost all libraries around the world have their systems convertible to the LCSH. The LCSH system is also superior to other world knowledge taxonomies. Table I presents a comparison of the LCSH with the Library of Congress Classification (LCC), the Dewey Decimal Classification (DDC), and the Yahoo! categorisation (YC). The LCSH has the largest number of topics, and the most specific semantic relations and structure. LCSH descriptors are classified by professionals, and the classification quality is guaranteed by well-defined and continuously-refined cataloging rules - in many respects, the LCSH has become a de facto standard for subject cataloging and indexing [9]. A world ontology constructed from the LCSH has also been proven promising by Tao et al. [35], for the problem of user background knowledge discovery from user local text documents. In summary, the LCSH is an ideal, competent world knowledge ontology for semantic categorisation of text documents.

The concepts in the world ontology are called *subjects* that are encoded from subject headings in the LCSH authorities. The semantic relationships of subjects are encoded from the references defined in the LCSH authorities for subject headings, such as *Broader Term*, *Used for*, and *Related to*. The $ancestor(s)$ function in Definition 2 returns the *Broader Term* subjects of s (they are semantically broader and thus, more general than s); the $descendant(s)$ returns the subjects that are *Used for* s and the subjects for which s is their *Broader Term* (s is semantically broader and thus, more specific than these subjects); the $neighbour(s)$ returns the *Related to* subjects of s.

B. Feature Discovery from Text

Given a document $d = \{t_1, t_2, \ldots, t_m\}$, let $\mathcal{F}(d) = \{\langle p, w(p)\rangle\}$, a set of competent patterns with weights, be the feature set of d. $\mathcal{F}(d)$ is to be discovered using the closed frequent sequential pattern mining technique.

We first introduce the concept of *sequential patterns*. A sequential pattern $p = \langle t_1, \ldots, t_r\rangle$ is an ordered list of terms. Given two sequential patterns p_1 and p_2, if p_1 is a subsequence of p_2, we say p_1 is a sub-pattern of p_2, and p_2 a super-pattern of p_1.

A pattern's *frequent* level depends on its occurrence frequency in the document. Let $P(d)$ be the set of all n-gram $(0 < n <= |d|)$ patterns that can be extracted from d; $termset(p)$ be a function that returns the set of terms in a pattern p and $termset(p) \subseteq d$. $coverset(p)$ is the covering set of p for d, and includes all patterns $p' \in P(d)$ satisfying $termset(p) \subseteq termset(p')$; $coverset(p) = \{p' | p' \in P(d), termset(p) \subseteq termset(p')\} \subset P(d)$. The absolute support $sup_a(p)$ is the number of occurrences of p in $P(d)$; $sup_a(p) = |coverset(p)|$. The relative support $sup_r(p)$ is the fraction of the patterns that contain $termset(p)$; $sup_r(p) = \frac{|coverset(p)|}{|P(d)|}$. p is then called *frequent pattern* if its sup_a (or sup_r) $\geq min\_sup$, a minimum support.

TABLE II
FEATURE DISCOVERED FROM THE SAMPLE DOCUMENT

Features	**Frequency**
dimens	2
espionag	4
econom espionag	3
secret	2
econom	4

We then define the concept of *closed* patterns. Given a set of patterns $P' \subseteq P(d)$, we can also define its *termset* by:

$$termset(P') = \{t | \forall p \in P' \Rightarrow t \in p\} \qquad (2)$$

The closure of a pattern p is defined as:

$$Cls(p) = termset(coverset(p)) \qquad (3)$$

A pattern p is then called *closed* if and only if $termset(p) = Cls(p)$.

The definition of *closed frequent sequential patterns* relies on a property of closed patterns. Given a closed pattern p, for all patterns $p_1 \supset p$, we have

$$sup_a(p_1) < sup_a(p) \qquad (4)$$

A frequent sequential pattern p is called *closed* if there exists no super-pattern p_1 of p such that $sup_a(p_1) = sup_a(p)$.

Based on these definitions, given a d, its feature set $\mathcal{F}(d) = \{\langle p, w(p)\rangle\}$ is discovered, where $w(p)$ is the frequency of p in d. Table II shows the features (closed frequent sequential patterns) discovered from the sample document illustrated in Fig. 2. Note that $min\_sup = 2$, and the feature discovery is based on the text following pre-processing.

C. Subject Extraction from World Ontology

Let $\mathbb{T}$ be the term space of $\mathbb{S}$ in $\mathcal{O}$ and $\mathbb{T} = \bigcup_{s \in \mathbb{S}} label(s)$. A matrix coordinated by $\mathbb{T}$ and $\mathbb{S}$ can be obtained:

***Definition** 7:* Let $\langle \mathbb{S}, \mathbb{T}\rangle$ be the matrix coordinated by $\mathbb{T}$ and $\mathbb{S}$, where a mapping exists:

$$\mu : \mathbb{T} \rightarrow 2^{\mathbb{S}}, \quad \mu(t) = \{s \in \mathbb{S} | t \in label(s)\} \subseteq \mathbb{S}$$

and its reverse mapping also exists:

$$\mu^{-1} : \mathbb{S} \rightarrow 2^{\mathbb{T}}, \quad \mu^{-1}(s) = \{t \in \mathbb{T} | s \in \eta(t)\} \subseteq \mathbb{T}.\square$$

By $\mu : \mathbb{T} \rightarrow 2^{\mathbb{S}}$, a term $t \in \mathbb{T}$ maps to a set of subjects $\mathcal{S}_t \subseteq \mathbb{S}$. Thus, given the feature set $\mathcal{F}(d) = \{\langle p, w(p)\rangle\}$, a set of subjects can be extracted from $\mathbb{S}$:

$$\mathcal{S}_d = \bigcup_{t \in termset(\mathcal{F}(d))} \mu(t) \qquad (5)$$

TABLE III
SUBJECTS EXTRACTED FOR THE SAMPLE DOCUMENT

Subject	Strength
Espionage	16.83
Espionage, economic	13.01
Space surveillance	13.01
Dimensions	9.24
Espionage, industry	9.24
Business espionage	8.98
Espionage literature	8.98
Espionage story	8.98
... ...	...

where $\mathcal{S}_d \subseteq \mathbb{S}$; $\mu(t) = \emptyset$ if $t \notin \mathbb{T}$.

By $\mu^{-1} : \mathbb{S} \to 2^{\mathbb{T}}$, a subject $s \in \mathbb{S}$ maps to a set of terms $\{t\} \subseteq \mathbb{T}$. Hence, with Eq. (5), a set of terms can be extracted from $\mu^{-1}(s)$ to expand d:

$$termset(d) = \bigcup_{s \in \mathcal{S}_d} \mu^{-1}(s) \tag{6}$$

Note that $termset(d) \neq d$. There exist some terms $\{t|t \in termset(d), t \notin d\}$ that are suggested by $\mathcal{S}_d$; there also exist some terms $\{t|t \notin termset(d), t \in d\}$ not in the term space $\mathbb{T}$ and thus, mapping to an empty subject set.

Because $\mathcal{S}_d$ is extracted using $\mathcal{F}(d) = \{\langle p, w(p)\rangle\}$, considering the weights of feature patterns, we can evaluate the terms $t \in termset(d)$:

$$w(t) = \sum_{p \in \{p|t \in termset(p), p \in \mathcal{F}(d)\}} w(p) \tag{7}$$

Considering the distribution of the terms spreading in other subject labels, the normalized form of term evaluation is defined as:

$$nw(t) = w(t) \times log(\frac{|\mathcal{S}_d|}{sf(t, \mathcal{S}_d)}) \tag{8}$$

where $sf(t, \mathcal{S}_d) = |\{s|t \in \mu^{-1}(s), s \in \mathcal{S}_d\}|$.

Subjects in $\mathcal{S}_d$ can finally be evaluated for their competence of summarizing d, using $nw(t)$ for all $t \in \mu^{-1}(s)$:

$$str(d, s) = \sum_{t \in \mu^{-1}(s)} nw(t) \tag{9}$$

By using the normalized form of terms, the subjects are competent for not only describing d but also distinguishing d from other documents in the document space Ω.

To prune away noisy subjects, a threshold, min_str, is applied to subject extraction. The subjects with $str(d, s) \geq min\_str$ are kept, whereas those with $str(d, s) < min\_str$ are dropped. During the experiments, different values were tested for min_str. The results revealed that setting min_str as the top 5th $str(d, s)$ value, a variable but a static value, gave the system the best performance. Table III shows the valid subjects extracted from the world ontology for the sample document in Fig. 2, using the features shown in Table II. Note that only the top subjects are displayed, because a total of 80 subjects survived the pruning process.

D. Generalising Subjects for categorisation

The subject set extracted from the ontology (as described in Section V-C) suffers from problems, such as the set being easily oversized and many subjects overlapping in their referring-to semantic space. As a result, the system complexity becomes high and its performance becomes difficult to handle when using the subject set. The extracted subject set must be generalized for semantic categorisation.

An example of how subjects extracted from the ontology overlap in their semantic space is displayed in Table III. Through common sense, we know that *Espionage* dominates *Espionage, economic*, *Espionage, industrial*, and *Business espionage*; that *Espionage literature* dominates *Espionage story*. This overlapping is caused by the same feature terms occurring in different subject labels. The overlapping space needs to be clarified and the noisy subjects need to be removed.

The algorithm of generalizing subjects is proposed based on the observation of the semantic overlapping of subjects. The algorithm is accomplished via investigating the relationships existing between these subjects. From Definitions 2 and 3, we know subjects in the world ontology are linked by semantic relations. Within the taxonomical structure, let s_1 and s_2 be two subjects and $s_1 \in ancestor(s_2)$ ($s_2 \in descendant(s_1)$). s_1 refers to a larger semantical extent than s_2, and thus, is more general than s_2. On the other hand, s_2 is more specific than s_1, thus focuses more on its referring-to topic. Such semantic relations can be revealed from an example. Let s_1 be *Automobile* and s_2 *Sedan*. *Automobile* contains *Car*, *Truck*, etc; *Car* contains *Sedan*, *Hatchback*, etc. *Automobile* covers broader extent than *Sedan*; vice versa, *Sedan* is more focused than *Automobile*. Therefore, if one subject is a descendant of another, the descendant can be removed because its referring-to semantical extent has already been covered by the other. By doing so, we have no information loss but limited focus (e.g., replacing *Sedan* by *Car*). With the same rule, if *Sedan* and *Hatchback* are both in the set, they may be replaced by their common ancestor *Car* without information loss, even if *Car* is not in the extracted set. Based on these, if some extracted subjects are under the same umbrella of an ancestor, their referring-to semantic extent is covered by that referred-to by their ancestor. Therefore, by losing no information but only limiting focus, we can replace these subjects with their ancestor, whether this common ancestor is in the extracted subject set or not.

The issue becomes how much focus we can afford to lose. A common ancestor chosen to replace its descendant subjects cannot be too far from the replaced descendants in the taxonomic structure, or the main focus will be lost. One extreme example is that we should never use *Thing* to replace any subject. *Thing* as the root dominates all subjects in the ontology. An ancestor subject being too far from its descendants reduces meaning. Therefore, we use only the lowest common ancestor (LCA) to replace the descendant subjects. The LCA is defined as the common ancestor of a set of subjects with the shortest distance to these subjects in the taxonomic structure of ontology. The LCA dominates descendant subjects and covers their semantic extent with only limited loss of focus.

```
input : $\mathcal{S}_i = \{s_1, s_2, \ldots, s_j\}$ (subject set extracted $i$), $\mathcal{O}$;
output: $\mathcal{S'}_i = \{s_1, s_2, \ldots, s_k\}$ (subject set generalized to map $i$).
$\mathcal{S'}_i = \emptyset, \mathcal{S}_{temp} = \emptyset, \mathcal{S}_{redundant} = \emptyset$;
foreach $s \in \mathcal{S}_i$ do
    Extract $S(s)$ from $\mathcal{O}$ where
      $S(s) = \{s' | s' \in ancestor(s), \delta(s \mapsto s') \leq 3\}$;foreach
      $s_n \in \mathcal{S}_i$ where $s_n \neq s$ do
        Extract $S(s_n)$ from $\mathcal{O}$ like Step 3;
        if $S(s) \cap S(s_n) \neq \emptyset$ then $\{\widehat{s} =$
          $\mathcal{LCA}(S(s) \cup S(s_n)), str(i, \widehat{s}) = str(i, s) + str(i, s_n)$;
          $\mathcal{S}_{temp} = \mathcal{S}_{temp} \cup \{\widehat{s}\}$;
          $\mathcal{S}_{redundant} = \mathcal{S}_{redundant} \cup \{s, s_n\}$;
        }
    end
    if $\mathcal{S}_{temp} \neq \emptyset$ then $\{\mathcal{S'}_i = \mathcal{S'}_i \cup \mathcal{S}_{temp}$;
      $\mathcal{S}_i = \mathcal{S}_i - \mathcal{S}_{redundant}$; $\mathcal{S}_{temp} = \emptyset$; $\mathcal{S}_{redundant} = \emptyset\}$;
    else $\mathcal{S'}_i = \mathcal{S'}_i \cup \{s\}$;
end
return $\mathcal{S'}_i$.
```

Algorithm 1: Generalizing Subjects

Algorithm 1 explains the process of semantic categorisation of a document via generalising the subjects initially extracted from the ontology. $\delta(s_1 \mapsto s_2)$ is a function measuring the distance between two subjects, which is completed by counting the number of edges travelled from s_1 to s_2 in the taxonomic structure of ontology. $\mathcal{LCA}(S(s_1) \cup S(s_2))$ is a function returning $\widehat{s}$, the LCA of s_1 and s_2 in a joint subject set, $S(s_1) \cup S(s_2)$.

Table IV presents the categorisation results generalized from the subjects displayed in Table III, with the min_str set as the top 5th $str(i, s)$ value again. Similar subjects like *Espionage*, *Espionage, economic*, *Espionage, industrial*, and *Business espionage*, have been merged and replaced by their LCA *Espionage* and *Business Intelligence*; *Espionage literature* and *Espionage story* replaced by *Spy story*. Consequently, the 80 subjects initially extracted from the world ontology (as described in Section V-C previously) are generalized to a much shorter list with only five subjects. This semantic categorisation result is meaningful, and in terms of semantics very close to the subjects listed with the sample document in Fig. 2, which were manually assigned by linguists.

VI. Experimental Evaluation

A. Experiment Design

Ideally, to categorise a document, the subjects automatically generated by the proposed approach should be exactly the same as those specified by specialist librarians. Though such a goal is unrealistic, the ideal scenario inspirited the design of our evaluation experiments. The proposed method was evaluated, based on the ground truth of manual assignment of subjects from linguists and compared against typical baseline classification methods.

TABLE IV
Generalized Subjects for the Sample Document

Subject	Strength
Espionage	269.78
Business Intelligence	203.83
Space surveillance	17.96
Spy story	16.27
Dimensions	9.24

TABLE V
Statistics of the Testing Set

Description	Stat.
Number of documents crawled	227,219
Number of documents used in experiments	31,902
Shortest length of documents in experiments	30
Longest length of documents in experiments	952
Average length of documents in experiments	85

The experiments were performed using a large testing set crawled from the catalogue of the University of Melbourne library[4]. The subject headings assigned to the catalogue items were manually specified by LCSH authorities through specialist librarians trained to specify subjects for a document without bias [9]. A sample catalogue item was presented in Section III. The title and content of catalogue items were used to form the content.

The text of each item in the catalogue was parsed first to remove unused information in this work, such as author name and Dewey Decimal Codes (Fig. 2 is an example document at this stage). The title and body of documents were equally removed during this process. General pre-processing techniques such as stopword removal and Porter stemming were applied to the preparation of the testing set for the experiment. Table V shows the statistics of the testing set (The length of documents refers to the number of terms in the documents after stopword removal). In the experiments, we used only documents having at least 30 terms. Documents shorter than that did not provide substantial frequent patterns, as revealed in preliminary experiments. By using the catalogue items in a library as the corpus, we could easily obtain a large testing set as well as a perfect ground truth for evaluation.

The subjects manually assigned to the documents by linguists provided the ideal ground truth in the experiments to measure the effectiveness of the proposed approach, against the automatically generated subjects. The objective evaluation methodology also assured the solidity and reliability of the experimental evaluation for our proposed method.

B. Baseline Models

Given that the LCSH ontology contains 394,070 subjects in our implementation, the semantic categorisation problem could also be understood as a $\mathcal{X}$-class classification problem where $\mathcal{X} = |\mathbb{S}| = 394,070$. Hence, we chose two typical multi-class classification approaches, *Rocchio* and kNN, for the baseline models in the experiments.

Rocchio is a simple and efficient classification method using centroid to define the class boundaries. The centroid of a subject s is computed as the vector average:

$$\vec{\mu}(s) = \frac{1}{|D_s|} \sum_{d \in D_s} \vec{v}(d)$$

[4] http://www.library.unimelb.edu.au/

In the experiments, the training set D_s contained only a single document $d = label(s)$. The $\vec{v}(d)$ was evaluated by using the frequency of terms in $label(s)$. The distance between a document and a subject class was measured by cosine similarity. The document was then classified into the subject classes with the top cosine value (Considering that $\mathcal{X} = |\mathbb{S}| = 394070$ is a huge number, using only the top value has already generated a considerably large set of subjects).

Unlike *Rocchio*, k Nearest Neighbour (kNN) determines the decision boundary locally and classifies documents into the major class of its k closest neighbours. When inputting a document d from the testing set, we extracted the closest neighbours $NN(d)$ that had the highest cosine similarity value with d. Because the testing documents were usually short, a large number of documents had the same cosine values. Thus, we set $k = 1$ to limit the number of considerable neighbours and ensure the highest possible accuracy. The distance of a s and a d is then evaluated by aggregating the cosine value of each $d' \in NN(d)$ to s. Again, d was classified into the subjects with only the top cosine value.

C. Performance Measuring Methods

The performance of the experimental models were measured by standard methods, precision and recall [5]. For the semantic categorisation problem, precision measured the ability of a method to categorise a document with highly-focused subjects, and recall with high-coverage of possible subjects.

As discussed previously, considering that $\mathcal{X} = |\mathbb{S}| = 394070$, pursuing the exact same subjects as those manually assigned by linguists is an unrealistic task. Thus, in respect to the testing set and the ground truth featured by the LCSH, performance was evaluated by:

$$precision = \frac{|\mathcal{FT}(S_{tgt}) \cap \mathcal{FT}(S_{grt})|}{|\mathcal{FT}(S_{tgt})|}$$

$$recall = \frac{|\mathcal{FT}(S_{tgt}) \cap \mathcal{FT}(S_{grt})|}{|\mathcal{FT}(S_{grt})|}$$

where $\mathcal{FT}(S) = \bigcup_{s \in S} \mu^{-1}(s)$ (see Definition 7); tgt referred to the target experimental model; grt referred to ground truth subjects.

In the experiments we also employed *micro-*F_1 Measure:

$$F_1 = \frac{2 \times precision \times recall}{precision + recall}$$

Precision and recall were evenly weighted in F_1 Measure. Each document's categorisation result was evaluated first and then all results were averaged for the final F_1 value. As with precision and recall, greater F_1 values indicated better performance.

VII. Results and Discussions

A. Experimental Results

Calling the proposed semantic categorisation approach the *OntoSum* model, the experiments compare the effectiveness of the performance of *OntoSum* against the baselines *Rocchio* and kNN models. Their effectiveness performances are depicted in Fig. 4, 5, and 6 for precision, recall, and F-Measure results, respectively. The value axis indicates the effectiveness rate between 0 and 1; the category axis indicates the number of documents whose categorisation results meet the indicating effectiveness rate. The number of documents is counted for those with only valid values (> 0).

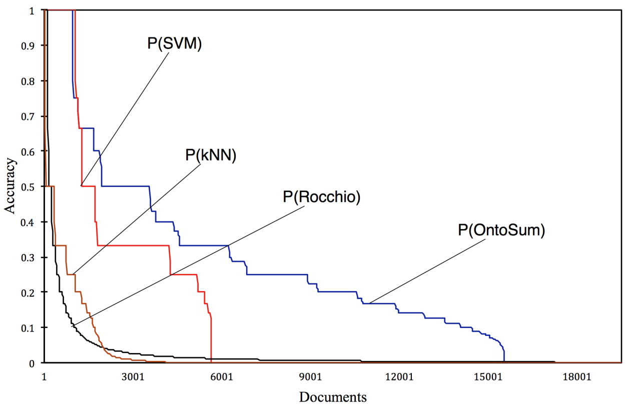


Fig. 4. Experimental Precision Results

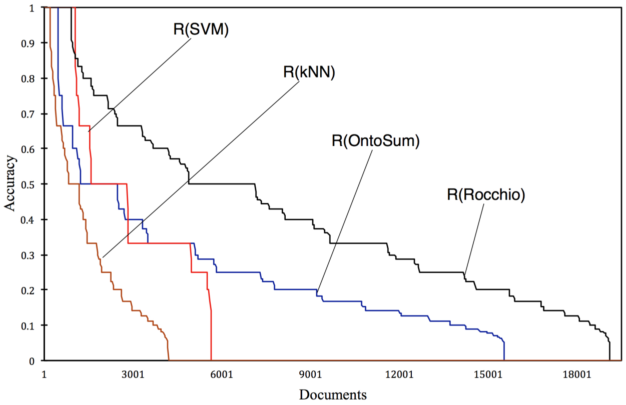


Fig. 5. Experimental Recall Results

The overall average performance is presented in Table VI. The F_1 measure equally considers both precision and recall when measuring performance. Thus the F_1 results are an overall effectiveness performance. The average F_1 results shown in Table VI reveal that the *OntoSum* model has achieved much better overall performance (0.125115) than the baseline models (0.019980 and 0.016305). This is also depicted in Fig. 6, where the $F(OntoSum)$ line is located at much higher bound level compared with the $F(Rocchio)$ and $F(kNN)$ lines.

Precision measures the accuracy of categorisation. For this, the *OntoSum* model also outperformed the baseline models. The average precision results in Table VI show this, with *OntoSum* 0.157992 vs. *Rocchio* 0.020259 and kNN 0.02077. Additionally, in Fig. 4, $P(OntoSum)$ is much higher than the

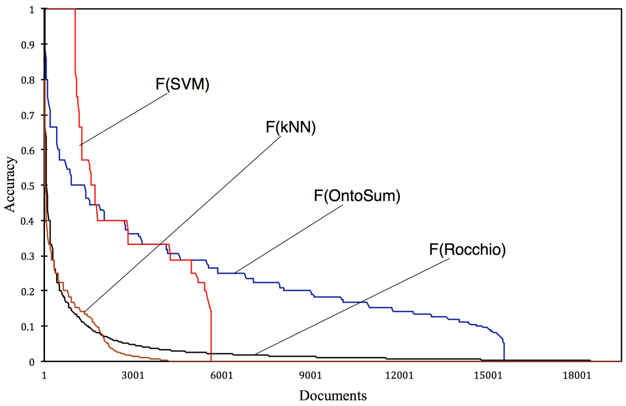


Fig. 6. Experimental F-Measure Results

TABLE VI
EFFECTIVENESS PERFORMANCE ON AVERAGE

	Precision	Recall	F-Measure
OntoMap	0.157992	0.134965	0.125115
SVM	0.0834775	0.093606	0.087678
Rocchio	0.020259	0.290226	0.019980
kNN	0.02077	0.053931	0.016305

other two.

Recall measures the semantic coverage of categorisation. The recall performance in the experiments shows a slightly different result compared to F_1 Measure and precision performance. The *Rocchio* model achieved the best recall performance (0.290226 on average), outperforming both the *OntoSum* (0.134965) and kNN model (0.053931). This is also illustrated in Fig. 5, in which $R(OntoSum)$ is in the middle of $R(Rocchio)$ and $R(kNN)$.

B. Discussions

There was a gap between the recall performance of the *OntoSum* and the baselines. After investigation, we found that the categorisation result of the *Rocchio* model was usually a large set of subjects (935 on average for each document), whereas the *OntoSum* model was 10 and the kNN 106. Due to the nature of recall, more features would be covered if the subject size became larger. As a result, the *Rocchio* categorisation with the largest size achieved the best recall performance. The subject sets generated by the kNN model had a larger size than those of the *OntoSum*. However, when taking neighbours into account, a large deal of noisy data was also brought into the neighbourhood - the average number of neighbours was 336. This was caused by the very large subject set in ontology and short documents. Thus, the categorisation became inaccurate, although only the subjects with the top similarity values were chosen to categorise a document. That is why the *OntoSum* sat in the middle of the *Rocchio* and kNN.

TABLE VII
SENSITIVITY STUDY RESULTS FOR TRACING A RIGHT NUMBER OF LEVELS IN ONTOLOGY TO FIND THE LOWEST COMMON ANCESTORS (LCAS)

	Precision	Recall	F-Measure
Level = 3	0.157992	0.134965	0.125115
Level = 5	0.154302	0.111632	0.111373

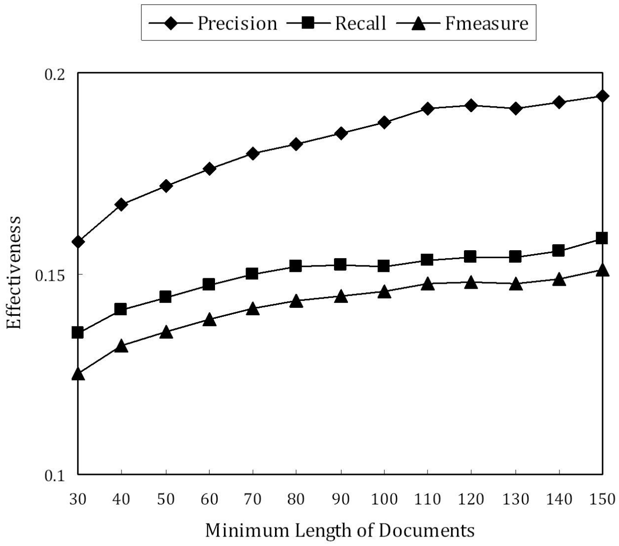


Fig. 7. Effectiveness of categorising Documents with Different Length

A different number of levels were tested in the sensitivity study for choosing the right number of levels to find the lowest common ancestor when generalising subjects for final categorisation (The relevant discussion is in Section V-D). Table VII displays the results for finding such a level. In the same experimental environment, when tracing three levels to find a LCA, the *OntoSum* model's performance - including F_1 Measure, precision, and recall - was better than that by five levels. In addition, tracing only three levels gave us lower complexity. Therefore, we chose three levels to find LCAs.

We also found that the performance of the *OntoSum* model slightly improved when the documents were relatively long. Figure 7 depicts the performance made by the *OntoSum* model on the documents with different minimum lengths. When the length of documents increased, the effectiveness sightly increased as well. Such an improvement is believed to be the result of the contribution of closed frequent sequential patterns discovered from documents (see Section V-B for details). When the *OntoSum* had the best performance with only documents longer than 150 terms, the average number of closed frequent sequential patterns was 27; when only with documents with length>= 90, the average number of patterns was 17; when considering all documents (length>= 30), the average number of discovered patterns dropped to 11. These results reveal that more useful and meaningful patterns would help the semantic categorisation in our approach. Given that more patterns would lead to more subjects extracted from the ontology, these facts also suggest that the generalising algorithm in the proposed approach successfully handled the extracted subjects well without sacrificing much information.

VIII. Conclusions

Semantic categorisation of text documents has become more important than ever, given that information in electronic form grown explosively. Many categorisation techniques have bottlenecks, such as being too expensive because of the large involvement of human effort, or are ineffective due to inadequate knowledge bases. The contribution of the work presented here addresses these bottlenecks, by introducing a semantic categorisation approach using a large world knowledge ontology built from the LCSH. A subject generalization algorithm has also been proposed in the work aiming to improve the performance of semantic categorisation. The approach was successfully evaluated through comparing typical text classification methods across a large testing set, measured by the categorisation manually made by linguists. This work contributes to text classification by demonstrating the value of an adequate and competent world knowledge ontology.

References

[1] B. Altinel, M.C. Ganiz, Semantic text classification: A survey of past and recent advances, Information Processing and Management, 54 (6) (2018) 1129-1153. doi:https://doi.org/10.1016/j.ipm.2018.08.001.

[2] B. Altınel, M.C. Ganiz, A new hybrid semi-supervised algorithm for text classification with class-based semantics, Knowledge-Based Systems, 108, 2016, 50–64.

[3] R. Bekkerman, M. Gavish, High-precision phrase-based document classification on a modern scale, in: Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, KDD '11, ACM, New York, NY, USA, 2011, pp. 231–239. doi:http://doi.acm.org/10.1145/2020408.2020449.
URL http://doi.acm.org/10.1145/2020408.2020449

[4] A. Bossard, Using document structure for automatic summarization, in: SIGIR '09: Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval, ACM, New York, NY, USA, 2009, pp. 849–849. doi:http://doi.acm.org/10.1145/1571941.1572170.

[5] C. Buckley, E. M. Voorhees, Evaluating evaluation measure stability, in: Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval, 2000, pp. 33–40.

[6] D. Cai, C. Zhang, X. He, Unsupervised feature selection for multi-cluster data, in: Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, KDD '10, ACM, New York, NY, USA, 2010, pp. 333–342. doi:http://doi.acm.org/10.1145/1835804.1835848.
URL http://doi.acm.org/10.1145/1835804.1835848

[7] F. Camastra, A. Ciaramella, A. Maratea, L.H. Son, A. Staiano, Semantic maps for knowledge management of web and social information, in: Computational Intelligence for Semantic Knowledge Management 2020 (pp. 39-51). Springer, Cham.

[8] F. Camous, S. Blott, A. F. Smeaton, Ontology-based medline document classification, in: Proceedings of the 1st international conference on Bioinformatics research and development, BIRD'07, Springer-Verlag, Berlin, Heidelberg, 2007, pp. 439–452.

[9] L. M. Chan, Library of Congress Subject Headings: Principle and Application, Libraries Unlimited, 2005.

[10] Z. Dou, R. Song, J.-R. Wen, A large-scale evaluation and analysis of personalized search strategies, in: WWW '07: Proceedings of the 16th international conference on World Wide Web, ACM Press, New York, NY, USA, 2007, pp. 581–590. doi:http://doi.acm.org/10.1145/1242572.1242651.

[11] E. Gabrilovich, S. Markovitch, Feature generation for text categorization using world knowledge, in: Proceedings of The Nineteenth International Joint Conference for Artificial Intelligence, Edinburgh, Scotland, 2005, pp. 1048–1053.
URL http://dl.acm.org/citation.cfm?id=1762370.1762413

[12] Y. Gao, Y. Xu, Y. Li, Pattern-based topics for document modelling in information filtering, IEEE Trans. Knowl. Data Eng. 27, 6 (2015), 1629—1642

[13] J. Gope, S.K. Jain, A survey on solving cold start problem in recommender systems, in: 2017 International Conference on Computing, Communication and Automation (ICCCA), IEEE, 2017 May 5, pp. 133–138.

[14] X. Hu, X. Zhang, C. Lu, E. K. Park, X. Zhou, Exploiting wikipedia as external knowledge for document clustering, in: KDD '09: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, New York, NY, USA, 2009, pp. 389–396. doi:http://doi.acm.org/10.1145/1557019.1557066.

[15] S. Huang, W. Peng, J. Li, D. Lee, Sentiment and topic analysis on social media: a multi-task multi-label classification approach, in: Proceedings of the 5th annual ACM web science conference, 2013, pp. 172–181.

[16] M. E. Houle, N. Grira, A correlation-based model for unsupervised feature selection, in: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management, CIKM '07, ACM, New York, NY, USA, 2007, pp. 897–900. doi:http://doi.acm.org/10.1145/1321440.1321570.
URL http://doi.acm.org/10.1145/1321440.1321570

[17] T. Joachims, Text categorization with Support Vector Machines: learning with many relevant features, in: Proceedings of the 10th European conference on machine learning, no. 1398, Springer Verlag, Heidelberg, DE, Chemnitz, DE, 1998, pp. 137–142.
URL citeseer.ist.psu.edu/joachims97text.html

[18] G. Kasper, D. de Siqueira Braga, D.M. Lima Martins, B. Hellingrath, User profile acquisition: A comprehensive framework to support personal information agents, in: Proceedings of the 2017 IEEE Latin American Conference on Computational Intelligence (LA-CCI), 2017, pp. 1–6. IEEE.

[19] Z. Kastrati, A.S. Imran, S.Y. Yayilgan, The impact of deep learning on document classification using semantically rich representations, Information Processing and Management 56(5), 2019, 1618–1632.

[20] I. Katakis, G. Tsoumakas, I. Vlahavas, Multilabel text classification for automated tag suggestion, in: In: Proceedings of the ECML/PKDD-08 Workshop on Discovery Challenge, 2008.

[21] S. Kulkarni, A. Singh, G. Ramakrishnan, S. Chakrabarti, Collective annotation of wikipedia entities in web text, in: KDD '09: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, New York, NY, USA, 2009, pp. 457–466. doi:http://doi.acm.org/10.1145/1557019.1557073.

[22] Y. Li, N. Zhong, Mining Ontology for Automatically Acquiring Web User Information Needs, IEEE Transactions on Knowledge and Data Engineering 18(4) (2006) 554–568.

[23] Y. Li, A. Algarni, S.-T. Wu, Y. Xu, Mining negative relevance feedback for information filtering, in: Proceedings of the IEEE/WIC/ACM international conference on Web Intelligence, 2009, pp. 606–613.

[24] Y. Li, A. Algarni, N. Zhong, Mining positive and negative patterns for relevance feature discovery, in: Proceedings of 16th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2010, pp. 753–762.

[25] B. Liu, Y. Dai, X. Li, W. Lee, P. Yu, Building text classifiers using positive and unlabeled examples, in: Proceedings of the Third IEEE International Conference on Data Mining, ICDM2003, 2003, pp. 179–186.

[26] P. Luo, F. Lin, Y. Xiong, Y. Zhao, Z. Shi, Towards combining web classification and web information extraction: a case study, in: KDD '09: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, New York, NY, USA, 2009, pp. 1235–1244. doi:http://doi.acm.org/10.1145/1557019.1557152.

[27] B. Maleszka, A method for ontology-based user profile adaptation in personalized document retrieval systems, in: 2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC), IEEE, 2016, pp. 003187-003192.

[28] H. H. Malik, J. R. Kender, Classifying high-dimensional text and web data using very short patterns, in: Proceedings of the 2008 Eighth IEEE International Conference on Data Mining, IEEE Computer Society, Washington, DC, USA, 2008, pp. 923–928. doi:10.1109/ICDM.2008.139.
URL http://dl.acm.org/citation.cfm?id=1510528.1511336

[29] G. Qiu, K. Liu, J. Bu, C. Chen, Z. Kang, Quantify query ambiguity using odp metadata, in: SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, ACM Press, New York, NY, USA, 2007, pp. 697–698. doi:http://doi.acm.org/10.1145/1277741.1277864.

[30] D. Ravindran, S. Gauch, Exploiting hierarchical relationships in conceptual search, in: Proceedings of the 13th ACM international conference on Information and Knowledge Management, ACM Press, New York, USA, 2004, pp. 238–239. doi:http://doi.acm.org/10.1145/1031171.1031221.

[31] F. Sebastiani, Machine learning in automated text categorization, ACM Computing Surveys (CSUR) 34 (1) (2002) 1–47. doi:http://doi.acm.org/10.1145/505282.505283.

[32] A. Sieg, B. Mobasher, R. Burke, Learning ontology-based user profiles: A semantic approach to personalized web search, The IEEE Intelligent Informatics Bulletin 8(1) (2007) 7–18.

[33] A. Sieg, B. Mobasher, R. Burke, Web search personalization with ontological user profiles, in: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management, ACM, New York, NY, USA, 2007, pp. 525–534. doi:http://doi.acm.org/10.1145/1321440.1321515.

[34] A. Singh, A. Sharma, N. Dey, Semantics and agents oriented web personalization: state of the art, International Journal of Service Science, Management, Engineering, and Technology (IJSSMET), 6,(2), (2015), 35–49.

[35] X. Tao, Y. Li, N. Zhong, A personalized ontology model for web information gathering, IEEE Transactions on Knowledge and Data Engineering, IEEE computer Society Digital Library. IEEE Computer Society 23 (4) (2011) 496–511. doi:http://doi.ieeecomputersociety.org/10.1109/TKDE.2010.145.

[36] J. Teevan, C. Alvarado, M. S. Ackerman, D. R. Karger, The perfect search engine is not enough: a study of orienteering behavior in directed search, in: Proceedings of the SIGCHI conference on Human factors in computing systems, 2004, pp. 415–422. doi:http://doi.acm.org/10.1145/985692.985745.

[37] J. Teevan, S. T. Dumais, E. Horvitz, Personalizing search via automated analysis of interests and activities, in: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval, 2005, pp. 449–456. doi:http://doi.acm.org/10.1145/1076034.1076111.

[38] X. Wan, J. Yang, Multi-document summarization using cluster-based link analysis, in: SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, ACM, New York, NY, USA, 2008, pp. 299–306. doi:http://doi.acm.org/10.1145/1390334.1390386.

[39] S. Wang, G. Englebienne, S. Schlobach, Learning concept mappings from instance similarity, in: A. Sheth, S. Staab, M. Dean, M. Paolucci, D. Maynard, T. Finin, K. Thirunarayan (Eds.), The Semantic Web - ISWC 2008, Vol. 5318 of Lecture Notes in Computer Science, Springer Berlin / Heidelberg, 2008, pp. 339–355.

[40] Y. Yan, N. Okazaki, Y. Matsuo, Z. Yang, M. Ishizuka, Unsupervised relation extraction by mining wikipedia texts using information from the web, in: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2 - Volume 2, ACL '09, Association for Computational Linguistics, Stroudsburg, PA, USA, 2009, pp. 1021–1029.
URL http://dl.acm.org/citation.cfm?id=1690219.1690289

[41] B. Yang, J.-T. Sun, T. Wang, Z. Chen, Effective multi-label active learning for text classification, in: KDD '09: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, New York, NY, USA, 2009, pp. 917–926. doi:http://doi.acm.org/10.1145/1557019.1557119.

[42] T. Yang, R. Jin, A. K. Jain, Y. Zhou, W. Tong, Unsupervised transfer classification: application to text categorization, in: Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, KDD '10, ACM, New York, NY, USA, 2010, pp. 1159–1168. doi:http://doi.acm.org/10.1145/1835804.1835950.
URL http://doi.acm.org/10.1145/1835804.1835950

[43] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, E. Hovy, Hierarchical attention networks for document classification, in: Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies, San Diego, California, USA, 2016, Jun pp. 1480–1489.

[44] L. Yu, S. Wang, K. K. Lai, An integrated data preparation scheme for neural network data analysis, IEEE Transactions on Knowledge and Data Engineering 18 (2) (2006) 217–230. doi:http://dx.doi.org/10.1109/TKDE.2006.22.

[45] L. Zadeh, Web intelligence and world knowledge - the concept of Web IQ (WIQ), in: Processing NAFIPS '04, IEEE Annual Meeting of the Fuzzy Information, 2004., Vol. 1, 2004, pp. 1–3.

[46] D. Zha, C. Li. Multi-label dataless text classification with topic modeling, Knowledge and Information Systems, 61 (2019) 137-–160. https://doi.org/10.1007/s10115-018-1280-0